ENVIRONMETRICS
Environmetrics 2000; **12**:1–10

A mixture model for the probability distribution of rain rate

Neal Jeffries[1] and Ruth Pfeiffer[2]*†

[1] *National Institute of Neurological Disorders and Stroke, 7550 Wisconsin Ave., Federal Bldg/7c06, Bethesda, MD 20892, U.S.A.*
[2] *National Cancer Institute, 6120 Executive Blvd/EPS 8017, Rockville, MD 20852, U.S.A.*

SUMMARY

In this paper we present a logistic mixture model for rain rate, that is, a model where the regime probabilities are allowed to change over time and are modeled with a logistic regression structure. Such a model may be used as an alternative to simple mixture, threshold, or hidden Markov models. The maximum likelihood estimates for the model parameters are found using an EM algorithm and their asymptotic properties are stated. The model is fit to hourly measurements of rain rate that are part of the GATE dataset. The results are compared with results from a standard mixture model and from a single density model. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: mixtures; logistic regression; EM algorithm; rain rate

1. INTRODUCTION

It is a climatological observation that rain rate distributions change with the time of day due to effects of the daily, or diurnal, heating cycle. The diurnal variability of rainfall has been investigated and documented for a variety of data sources. A recent example is a paper by Soman *et al.* (1995), who averaged rain rate data over an area of the order of 10^5 km^2 from Darwin, Australia to obtain a time series. By fitting periodograms and correlograms, they found strong evidence for diurnal variation.

In addition to the variation of rain activity due to the diurnal effects of heating, many scientists have suggested there are different types of rain (see, for example, Houze, 1981) which indicates that a mixture density may be appropriate. Bell and Suhasini (1994), for example, used a non-parametric principal components approach to estimate a mixture of densities.

In this paper we model the distribution of rain rate data as a two regime mixture with the component distributions corresponding to stratiform (i.e. moderate), and convective (heavy),

* Correspondence to: R. Pfeiffer, National Cancer Institute, 6120 Executive Blvd/EPS 8017, Rockville, MD 20852, U.S.A.
† E-mail: pfeiffer@mail.nih.gov

Received 6 March 1999
Revised 12 January 2000

rain patterns. To incorporate the diurnal variation of the rain rate measurements, we let the mixing probabilities depend on time through a logistic regression.

The logistic mixture model is presented in detail in the next section. In Section 2 we also describe an EM algorithm approach to finding maximum likelihood estimates for the parameters in the two component distributions as well as those parameters in the logistic regression that predict the regime probabilities. In Section 3 we apply the model to rain rate data from the GATE dataset. As a basis of comparison we include results for two nested models. The first is a standard mixture model with constant regime probabilities, and the second model for comparison is a single component, or no-mixture, model. In the last section we discuss the fit of the different models and some of the related problems of identifiability that arise in the context of mixture models.

2. MODEL FORMULATIONS AND ESTIMATION METHOD

Mixture models were developed as a way of analyzing data that arise from two or more distinct data generation processes. Refer to McLachlan and Basford (1988), Titterington *et al.* (1985), or Everitt and Hand (1981) for good introductions to mixtures. The logistic mixture models we consider in this paper are characterized as follows. The data are pairs $(Y_j, z_j), j = 1, \ldots, n$, where Y_j denotes the measurement of interest and z_j is a $p \times 1$ vector of covariates associated with the jth measurement. We assume for brevity that there are only two states (or regimes), which we label as state 1 and state 0, described by the random variable I. (The extension of our analysis to more than two states is straightforward.) The state probabilities for the jth observation depend on the covariate vector z_j through a logistic regression model:

$$\mathbb{P}[I_j = 1 | z_j] = p(z_j; \gamma) = \frac{\exp(z'_j\gamma)}{1+\exp(z'_j\gamma)}. \tag{1}$$

The first component of z is equal to unity to allow for an intercept, and γ is the $p \times 1$ vector of associated, unknown logistic regression coefficients.

Given z, the probability density function of y is given by the mixture model

$$g(y|z, \theta) = f(y; \alpha_0) \cdot (1 - p(z; \gamma)) + f(y; \alpha_1) \cdot p(z; \gamma), \tag{2}$$

where $f(y; \alpha)$, $\alpha \in A \subseteq \mathbb{R}^l$, for some l, denotes a class of parametric density functions, and $\theta = (\alpha_0, \alpha_1, \gamma)$. We interpret $f(y; \alpha_k)$ to be the conditional density of y_j given $I_j = k$, for $k \in \{0, 1\}$.

2.1. *The estimation method*

We now discuss an EM (Expectation-Maximization) algorithm (see Dempster *et al.*, 1977; Wu, 1983; McLachlan and Krishnan, 1997) for mixtures of the type defined above. The EM algorithm is an alternative estimation procedure that is particularly well suited to approach problems with missing, or unobserved, data. In this instance, the unobserved data form the knowledge of which regime produced a given observation. The algorithm is used to maximize the likelihood, or equivalently, the log likelihood given by $\Sigma_{j=1}^{n} \log g(y_j|z_j; \theta)$, with g from Equation (2). It should

be noted that an EM approach is not necessary – the parameters could be estimated by software that has minimization or maximization routines (e.g. nlmin or nlminb functions in S-PLUS, or routines in FORTRAN or C). We want to point out though, that $\Sigma_j \log g(y_j|z_j;\theta)$ will not be a concave function because it is a mixture likelihood, and thus the Newton–Raphson based optimizations may not work well. The advantage of using an EM algorithm lies in the fact that the likelihood values increase (weakly) with each iteration. This means that if θ^m denotes the EM algorithms estimate of the true parameters after m iterations, then the EM estimates satisfy $L_n(\theta^{m+1}) \geqslant L_n(\theta^m)$. This monotone property of the EM estimates, that is clearly not true for estimates obtained by Newton–Raphson, becomes more important as the number of estimated parameters increases (either because the model may have more than two component densities, or the component densities are extended to have more parameters). As more parameters are added, it is more likely that standard maximization routines (e.g. Newton–Raphson based methods) will fail. There are cases in which the EM estimates may converge to a critical point other than a local maximum, or other difficulties may occur (see Wu, 1983; McLachlan and Basford, 1988), but such aberrations are usually overcome by changing the starting values of the algorithm. Wu (1983) states conditions that ensure that all limit points of any instance of the EM algorithm are local maximizers of $L_n(\theta)$. The implementation and interpretation of the EM algorithm in the context of mixture models has been discussed by several authors, including Titterington *et al.* (1985), McLachlan and Basford (1988), and McLachlan and Krishnan (1997). The derivation of the algorithm for logistic mixtures (i.e. variable probabilities of the form $\exp(z'\gamma)/(1+\exp(z'\gamma))$ is straightforward, detailed in Jeffries (1998), and will not be presented here. The procedure may be summarized as follows: given observed outcomes $y_1, y_2, \ldots, y_n$ and covariates $z_1, z_2, \ldots, z_n$,

1. Choose initial parameter values $\theta^1 = (\alpha_0^1, \alpha_1^1, \gamma^1)$.
2. Given $\theta^m = (\alpha_0^m, \alpha_1^m, \gamma^m)$ (the estimate after m iterations of the algorithm) calculate

$$\tilde{p}_j^m = \frac{f(y_j;\alpha_1^m)p(z_j;\gamma^m)}{f(y_j;\alpha_1^m)p(z_j;\gamma^m)+f(y_j;\alpha_0^m)(1-p(z_j;\gamma^m))} \quad \text{where} \quad p(z_j;\gamma) = \frac{\exp(z'_j\gamma)}{1+\exp(z'_j\gamma)}.$$

3. Given $\{\tilde{p}_j^m\}$ find α_0^{m+1}, α_1^{m+1}, and γ^{m+1} where

$$\alpha_1^{m+1} = \arg\max_{\alpha} \sum_j \tilde{p}_j^m \log f(y_j;\alpha) \tag{3}$$

$$\alpha_0^{m+1} = \arg\max_{\alpha} \sum_j (1-\tilde{p}_j^m) \log f(y_j;\alpha) \tag{4}$$

and

$$\gamma^{m+1} = \arg\max_{\gamma} \sum_j \tilde{p}_j^m \log p_j(\gamma) + (1-\tilde{p}_j^m)\log(1-p_j(\gamma)). \tag{5}$$

4. Repeat steps 2 and 3 until $\|\theta^{m+1}-\theta^m\|$ and $\Sigma_j \log g(y_j|z_j;\theta^{m+1}) - \Sigma_j \log g(y_j|z_j;\theta^m)$ are smaller than some prespecified tolerance level.

$$w = \frac{2\pi}{24}.$$

The functional form of p^s and p^c imposes a diurnal cycle as long as $\alpha \neq 0$. The α parameter controls the variability, or amplitude in the cycle. The β gives freedom to the phase shift of the cycle and the δ allows these probabilities to fluctuate about some average value different from 1/2. To put the regime probabilities for our rain rate model, $\exp(\alpha \sin(wh_j+\beta)+\delta)/(1+\exp(\alpha \sin(wh_j+\beta)+\delta))$, in the form $\exp(z_j'\gamma)/(1+exp(z_j'\gamma))$, we use that $\sin(x+y) = \cos x \sin y + \cos y \sin x$, and rewrite $\alpha \sin(wh_j+\beta)+\delta$ as $z_j'\gamma$, where $z_j = (1, \cos wh_j, \sin wh_j)$ and $\gamma = (\delta, \alpha \sin\beta, \alpha \cos\beta)$.

The probability density function of our logistic mixture model for the rain rate measurements is thus given by

$$g(x_j|h_j; f^c, f^s, \phi_c, \alpha, \beta, \delta)) = p^s(h_j; \alpha, \beta, \delta) f^s(x_j) + p^c(h_j; \alpha, \beta, \delta) f^c(x_j). \tag{8}$$

s and c are labels designating stratiform and convective and $f^c(\cdot)$ and $f^s(\cdot)$ denote the densities of rain rate for the two regimes. We additionally assume that the densities for convective and stratiform rain rates have the same functional form and differ only by their parameter values. Following the suggestions of Kedem *et al.* (1990, 1997), we take f to be the density of a two parameter lognormal distribution

$$f(x) = \begin{cases} 1/(\phi x\sqrt{2\pi})\exp(-(\log x - \mu)^2/2\phi^2) & x > 0, \\ 0 & x \leqslant 0. \end{cases}$$

Note that the logarithm of X has a normal distribution. We thus let $y_j = \log x_j$. The parametric model is now given by

$$g(\log x_j|h_j; \psi = (\mu_s, \phi_s, \mu_c, \phi_c, \alpha, \beta, \delta)) = g(y_j|h_j; \psi = (\mu_s, \phi_s, \mu_c, \phi_c, \alpha, \beta, \delta)) \tag{9}$$

$$= p^s(h_j; \alpha, \beta, \delta) f(y_j; \mu_s, \phi_s) + p^c(h_j; \alpha, \beta, \delta) f(y_j; \mu_c, \phi_c), \tag{10}$$

where f stands for the normal density with regime dependent mean μ_k and variance ϕ_k^2, for $k = s, c$.

As a basis of comparison we include results for two nested models:

$$g(y_j; \mu_s, \phi_s, \mu_c, \phi_c, p) = pf(y_j; \mu_s, \phi_s) + (1-p) f(y_j; \mu_c, \phi_c) \tag{11}$$

and

$$g(y_j; \mu, \phi) = f(y_j; \mu, \phi). \tag{12}$$

The model in (11) corresponds to a standard mixture model with fixed regime probabilities and the model in (12) is a one regime, or no-mixture model.

Table I. Three models of rain rate.

Parameter	LM	2R	1R
μ_s	−0.0930	−1.06	0·497
	(0.0831)	(0.0984)	(0.0436)
ϕ_s	0.907	0.4124	1.63
	(0.104)	(0.0934)	(1.81)
μ_c	1.74	0.709	NA
	(0.160)	(0.0944)	NA
ϕ_c	0.870	1.205	NA
	(0.157)	(0.0972)	NA
d	1.12	−1.994	NA
	(0.334)	(0.415)	NA
β	0.766	NA	NA
	(0.116)	NA	NA
α	1.48	NA	NA
	(0.234)	NA	NA
Log likelihood	−1380.97	−1417.05	−1430.09

In Table I we present point estimates and standard errors (in parentheses) for the three models. The column headed LM corresponds to results for logistic mixtures, the 2R heading denotes the mixture with constant regime probabilities in (11) and the 1R indicates results for the one regime model.

4. DISCUSSION

First we interpret the results of fitting the LM model. From the parameter estimates we obtain mean rain rates for each of the two densities. The mean for the stratiform regime is $\exp(-0.0930+0.5\cdot 0.907) = 1.43$ mm/hour and the mean for the convective regime is 8.82 mm/hour. From these means and the derived hourly regime probabilities, we produce estimated hourly rain rates. A plot of the stratiform regime probabilities and the estimated hourly rates is included in Figure 1. The plots indicate that the more intensive rain rates are associated with the afternoon.

The results of Table I indicate the surprising degree to which our estimates of stratiform and convective parameters differ between the LM and 2R model. Not only are the parameters of the component densities quite dissimilar, but the estimated regime probabilities are also markedly different. The unconditional estimate of the stratiform regime probability, $p^s = 0.678$, from LM is obtained as $\Sigma_h p^s(h;\alpha,\beta,\delta)P(h)$, where $P(h)$ is the observed proportion of the data at hour h. This is considerably higher than the proportion of the stratiform observations estimated from the standard mixture model, 0.12. This discrepancy may best be explained through examination of a histogram of the log rain rates (see Figure 2). Figure 2 also contains plots of the estimated densities of the LM and 2R regime, where the LM plot was created with the averaged $p^s = 0.678$. The histogram and density plots indicate the 2R log likelihood may be maximized around a different pair of modes than those that maximize the LM log likelihood. This finding suggests we should be diligent in making sure the log likelihood is maximized at the reported values. In our investigations the 2R and LM estimates were both robust to different starting values, as well

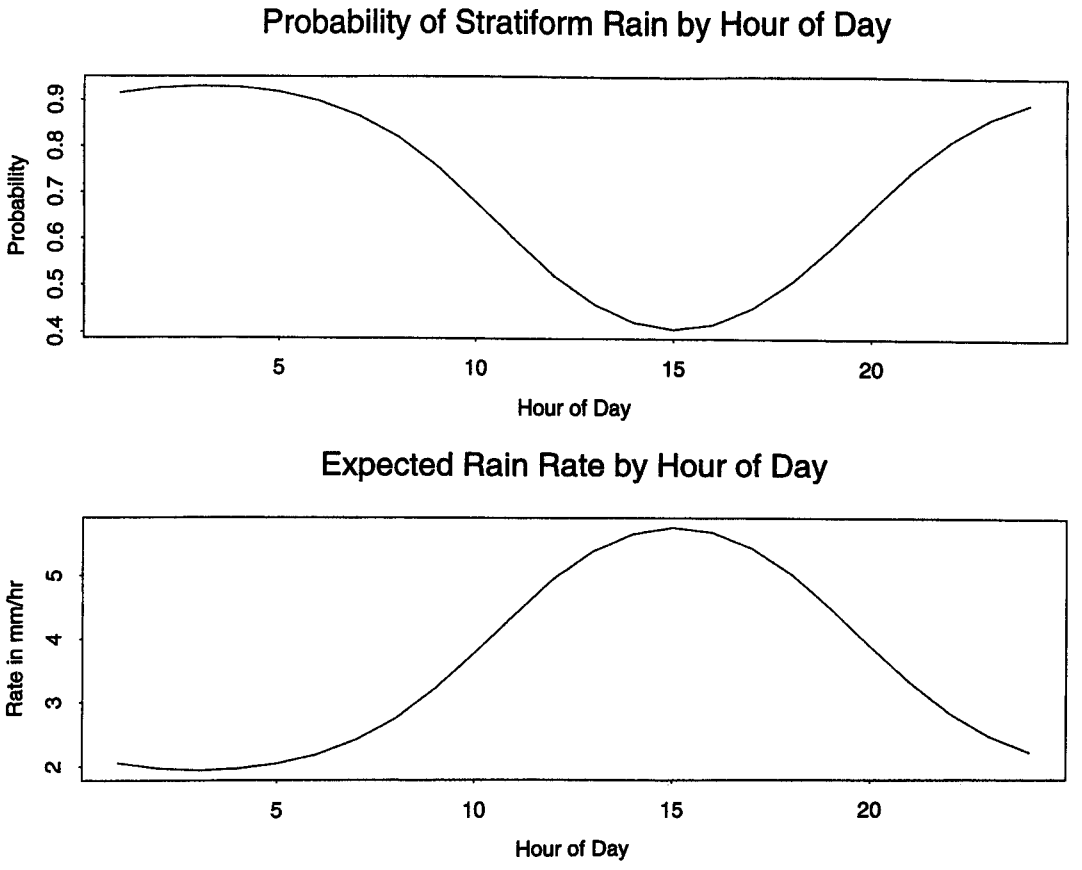

Figure 1. Estimated regime probabilities and rain rates.

as different optimization methods (both the EM algorithm approach outlined in Section 2, as well as generic maximization routines, were used). We are not sure how to explain why the two models pick out such different component densities other than to surmise that the parameterization of the diurnal cycle captures effects that are obscured when constant regime probabilities are imposed. These results also suggest that perhaps investigation of a three regime model is warranted. Bell and Suhasini (1994) used a non-parametric principal components approach to estimating a mixture of densities and found support for two distributions (with mean rain rates of 2.6 and 8.8 mm/hour – very similar to our results) but the data were not better fit by allowing for three distributions. We performed some estimation of three regime models with fixed probabilities, but the estimates were not robust (i.e. different starting points led to different estimates) and the greatest log likelihood we were able to achieve was -1410.5 still inferior to that obtained under the two regime LM model (-1381). The results also suggest that in passing to each of the more restrictive nested models the explanatory power is significantly weakened. We will discuss a formal test of such hypotheses in a subsequent paper. What may be most surprising is the apparent power that is gained by parameterizing the regime probabilities

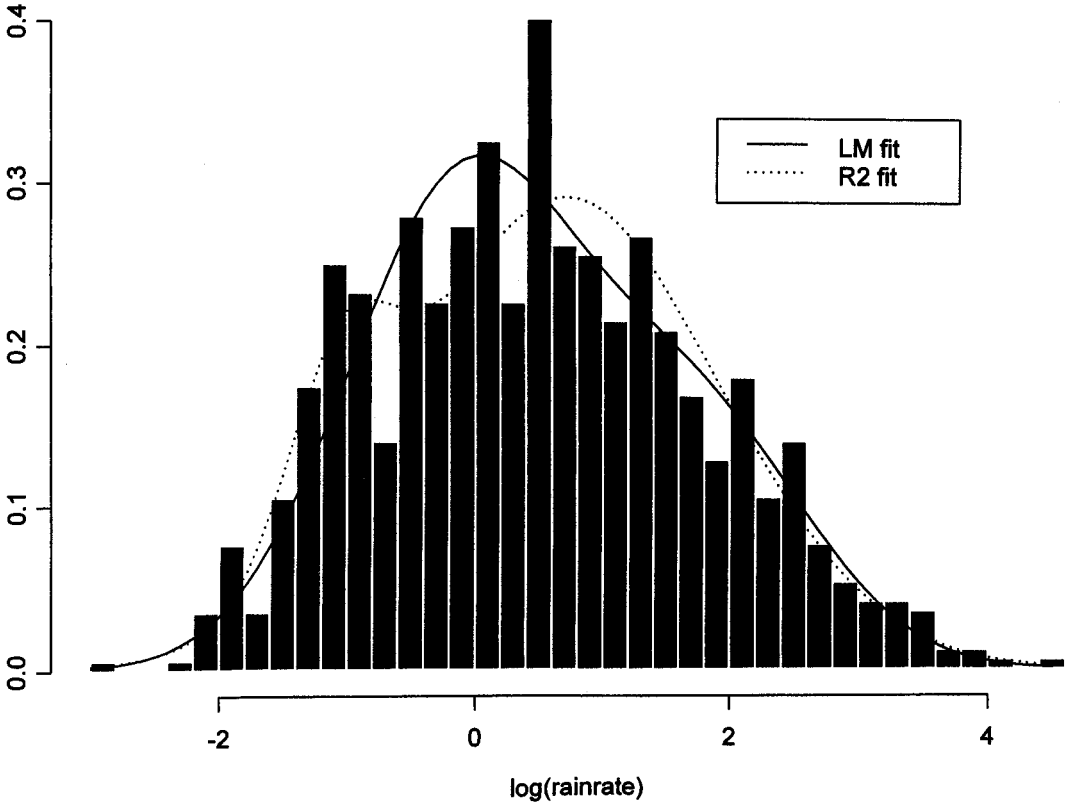

Figure 2. Estimated densities and histogram of log rain rates.

according to a daily cycle. The addition of the β and α parameters increases the log likelihood by approximately 36 over what was obtained from the 2R model. One might want to conclude that under the null hypothesis that the 2R model is correctly specified, then

$$2 * (\text{log likelihood(LM)} - (\text{log likelihood(2R)}) \xrightarrow{\mathscr{D}} \xi_2^2.$$

This would be incorrect as the β term is not identified under the null hypothesis that $p^s(h_j; \alpha, \beta, \delta) = p$, a constant. By this we mean, while it is clear that $\alpha = 0$ under the null hypothesis, any value of β would suffice, and hence β is not identified. A similar, though more difficult, identification problem arises if we want to compare either the LM or 2R model to the 1R model. These identifiability problems invalidate traditional approaches to hypothesis testing. Empirical process theory has been applied to yield asymptotic distributions of the likelihood ratio statistic in many instances – see Dacunha-Castelle and Gassiat (1997) for approaches to general mixtures and Jeffries (1998) for logistic mixtures in particular. Others (Feng and McCulloch, 1996; McLachlan *et al.*, 1993) advocate a bootstrap or Monte Carlo approach to testing whether the data are generated by a 1R or a 2R model (with obvious extensions for treating an LM model).

A full description of the ideas in these papers exceeds the scope of this discussion. We will take up these questions at length in future work.

ACKNOWLEDGEMENTS

We thank Dr Paul Smith for his helpful remarks and comments and Dr Tom Bell for bringing the rain application to our attention. N.J. is supported by the Patricia Roberts Harris Fellowship. R.P. is supported by NASA grant NGT-30332.

REFERENCES

Bell TL, Suhasini R. 1994. Principal modes of variation of rain-rate probabilitiy distributions. *Journal of Applied Meteorology* **33**:1067–1078.

Dacunha-Castelle D, Gassiat E. 1997. Testing in locally conic models, and application to mixture models. *ESIAM: Probability and Statistics* **1**:285–317.

Dempster DA, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**:1–22.

Everitt BS, Hand DJ. 1981. *Finite Mixture Distributions*. Chapman and Hall: London.

Feng ZD, McCulloch CE. 1996. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B* **58**:609–617.

Houze RA. 1981. Structures of atmospheric precipitation: a global survey. *Radio Science* **16**:671–689.

Hudlow MD, Patterson VL. 1979. *GATE Radar Rainfall Atlas*. NOAA Special Report, U.S. Government Printing Office: Washington.

Jeffries N. 1998. *Logistic Mixtures of Generalized Linear Model Time Series*. Ph.D. Thesis, University of Maryland.

Kedem B, Chiu LS, North GR. 1990. Estimation of mean rain rate: application to satellite observations. *Journal of Geophysical Research* **95**:1965–1972.

Kedem B, Pfeiffer R, Short DA. 1997. Variability of space-time mean rain rate. *Journal of Applied Meteorology* **36**:443–451.

McLachlan GJ, Basford KE. 1988. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker: New York.

McLachlan GJ, Krishnan T. 1997. *The EM Algorithm and Extensions*. Wiley: New York.

McLachlan GJ, Basford KE, Green M. 1993. On inferring the number of components in normal mixture models. Research Report #9, Department of Mathematics, The University of Queensland: Australia.

Soman VV, Valdes JB, North GR. 1995. Satellite sampling and the diurnal cycle statistics of Darwin rainfall data. *Journal of Applied Meteorology* **34**:2481–2490.

Titterington DM, Smith AFM, Makov UE. 1985. *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York.

Wu CF. 1983. On the convergence properties of the EM algorithm. *Annals of Statistics* **11**:95–103.

